Predicting bovine tuberculosis status of dairy cows from mid-infrared spectral data of milk using deep learning

Denholm, SJ; Brand, W; Mitchell, Andrew; Wells, AM; Krzyzelewski, T; Smith, SL; Wall, E; Coffey, MP

Published in:
Journal of Dairy Science

DOI:
10.3168/jds.2020-18328

Print publication: 01/10/2020

Document Version
Publisher's PDF, also known as Version of record

J. Dairy Sci. 103
https://doi.org/10.3168/jds.2020-18328

Predicting bovine tuberculosis status of dairy cows from mid-infrared spectral data of milk using deep learning

S. J. Denholm,[1]* W. Brand,[1] A. P. Mitchell,[2] A. T. Wells,[1] T. Krzyzelewski,[1] S. L. Smith,[1] E. Wall,[1] and M. P. Coffey[1]
[1]Scotland's Rural College (SRUC), Peter Wilson Building, Kings Buildings, West Mains Road, Edinburgh EH9 3JG, UK
[2]Animal and Plant Health Agency (APHA), Woodham Lane, Addlestone, Surrey KT15 3NB, UK

ABSTRACT

Bovine tuberculosis (bTB) is a zoonotic disease in cattle that is transmissible to humans, distributed worldwide, and considered endemic throughout much of England and Wales. Mid-infrared (MIR) analysis of milk is used routinely to predict fat and protein concentration, and is also a robust predictor of several other economically important traits including individual fatty acids and body energy. This study predicted bTB status of UK dairy cows using their MIR spectral profiles collected as part of routine milk recording. Bovine tuberculosis data were collected as part of the national bTB testing program for Scotland, England, and Wales; these data provided information from over 40,500 bTB herd breakdowns. Corresponding individual cow life–history data were also available and provided information on births, movements, and deaths of all cows in the study. Data relating to single intradermal comparative cervical tuberculin (SICCT) skin-test results, culture, slaughter status, and presence of lesions were combined to create a binary bTB phenotype labeled 0 to represent nonresponders (i.e., healthy cows) and 1 to represent responders (i.e., bTB-affected cows). Contemporaneous individual milk MIR spectral data were collected as part of monthly routine milk recording and matched to bTB status of individual animals on the single intradermal comparative cervical tuberculin test date (±15 d). Deep learning, a sub-branch of machine learning, was used to train artificial neural networks and develop a prediction pipeline for subsequent use in national herds as part of routine milk recording. Spectra were first converted to 53 × 20-pixel PNG images, then used to train a deep convolutional neural network. Deep convolutional neural networks resulted in a bTB prediction accuracy (i.e., the number of correct predictions divided by the total number of predictions) of 71% after training for 278 epochs. This was accompanied by both a low validation loss (0.71) and moderate sensitivity and specificity (0.79 and 0.65, respectively). To balance data in each class, additional training data were synthesized using the synthetic minority over sampling technique. Accuracy was further increased to 95% (after 295 epochs), with corresponding validation loss minimized (0.26), when synthesized data were included during training of the network. Sensitivity and specificity also saw a 1.22- and 1.45-fold increase to 0.96 and 0.94, respectively, when synthesized data were included during training. We believe this study to be the first of its kind to predict bTB status from milk MIR spectral data. We also believe it to be the first study to use milk MIR spectral data to predict a disease phenotype, and posit that the automated prediction of bTB status at routine milk recording could provide farmers with a robust tool that enables them to make early management decisions on potential reactor cows, and thus help slow the spread of bTB.

Key words: bovine tuberculosis, deep learning, mid-infrared spectroscopy, dairy cow, noninvasive

INTRODUCTION

Different physiological processes can leave molecular signatures in the milk of dairy cows (Soyeurt et al., 2006). Such signatures can potentially be detected by analyzing mid-infrared (**MIR**) spectral data, a byproduct resulting from routine milk recording, and used as biomarkers for economically important traits (Soyeurt et al., 2006, 2011). Mid-infrared spectroscopy of milk samples is an internationally used noninvasive method for the prediction of milk fat and protein content during routine milk recording. This method of prediction is increasingly being used as an efficient and effective low-cost tool for rapid prediction of expensive and, more often than not, difficult-to-record phenotypes. The utility of using milk MIR spectra as a phenotyping tool has become an increasingly popular area of research over the last 15+ years (Berry et al., 2013; De Marchi

Received February 7, 2020.
Accepted June 9, 2020.
*Corresponding author: scott.denholm@sruc.ac.uk

et al., 2014), with success demonstrated in the prediction of milk fatty acids (Soyeurt et al., 2011), body energy (McParland et al., 2011; Smith et al., 2019), methane emissions (Dehareng et al., 2012), ketone bodies (Grelet et al., 2016), lactoferrin (Soyeurt et al., 2012), feed intake (Wallén et al., 2018), and pregnancy status (Lainé et al., 2014; Toledo-Alvarado et al., 2018; Delhez et al., 2020). Further, such research has resulted in successful international and multidisciplinary collaborative projects such as RobustMilk (Veerkamp et al., 2013) and OptiMIR (Friedrichs et al., 2015). Moreover, for farmers already involved in routine milk recording, obtaining additional MIR spectra-based herd information requires no extra labor costs or changes in herd management. For milk-recording agencies, this data can be offered as an additional service to dairy farmers for only incremental data-handling costs.

Large data sets, such as those containing MIR spectral records, offer an exceptional opportunity to exploit the power of machine-learning algorithms to investigate and better understand relationships between milk spectra and traits of importance that may go otherwise unnoticed using other, or unsuitable, statistical techniques. Deep learning, a sub-branch of machine learning, uses algorithms and techniques that are better able to make use of the increasingly huge data sets and advances in computer technology of the present day (Bengio, 2009; Deng et al., 2009; Krizhevsky et al., 2012; LeCun et al., 2015).

Recently our group applied a deep convolutional neural network (**CNN**) to MIR-matched pregnancy data to predict the pregnancy status of dairy cows (Brand et al., 2018). We observed that milk MIR spectra contained features relating to pregnancy status and underlying metabolic changes in dairy cows, and that such features can be identified using a deep-learning approach. In our study, we defined pregnancy status as a binary trait (i.e., pregnant, not-pregnant) and found CNN significantly improved prediction accuracy, with trained models able to detect 83 and 73% of onsets and losses of pregnancy, respectively (Brand et al., 2018). More recently we have improved prediction accuracy such that models predict pregnancy status with an accuracy of 97% (with a corresponding validation loss of 0.08) after training for 200 epochs (our unpublished data).

Since proving the concept of training MIR spectra to predict a categorical (binary) trait using a deep-learning approach (i.e., pregnancy status in dairy cows), we have extended the technique to predict other hard-to-record phenotypes from MIR spectral data, specifically disease traits such as bovine tuberculosis (**bTB**).

Bovine tuberculosis is a zoonotic disease endemic in the UK and Ireland, and is distributed worldwide in parts of Africa, Asia, Europe, the Middle East, the Americas, and New Zealand (Humblet et al., 2009). This chronic, slowly progressive, and debilitating disease presents a significant challenge to the UK cattle sector and has considerable public health implications in countries where it is not subject to mandatory eradication programs (Olea-Popelka et al., 2017). The disease is caused *by Mycobacterium bovis* infection, primarily involving the upper- and lower-respiratory tracts and associated lymph nodes (Pollock and Neill, 2002). The Department for Environment, Food and Rural Affairs (Defra) lists bTB as one of the 4 most important livestock diseases globally, incurring annual costs of about £175 million in the UK ($227 million USD). In 2017, the total numbers of cows slaughtered due to bTB (i.e., all cows defined as reactors and inconclusive reactors) in England, Wales, and Scotland were 33,238, 10,053, and 273 cows, respectively, equating to a 14, 1, and 46% increase in the number of cows slaughtered compared with 2016 (Department for Environment, Food and Rural Affairs, 2018). The disease affects animal health and welfare, causing substantial financial strain on the dairy cattle sector worldwide through involuntary culling, animal movement restrictions, and the cost of control and eradication programs (Allen et al., 2010). Moreover, the disease has significant, and often unseen, social and psychological effects on farmers, particularly mental health (Parry et al., 2005; FarmingUK, 2018; Crimes and Enticott, 2019).

Recent research has led to the development of the world's first national genetic and genomic evaluation for bTB resistance in the Holstein dairy breed in the UK and the launch of the index TB Advantage (AHDB Dairy, 2016; Banos et al., 2017). Research confirmed the existence of significant genetic variation among individual animals for resistance to bTB infection, mainly inferred from the single intradermal comparative cervical tuberculin (**SICCT**) skin test and the presence of lesions and bacteriological tests following slaughter (Pollock and Neill, 2002; Bermingham et al., 2009; Brotherstone et al., 2010; Tsairidou et al., 2014). Initial research on dairy genetic evaluations for bTB has now been extended to all dairy breeds.

The objective of the present study was to use phenotypic reference data obtained from the Great Britain (**GB**) national bTB testing program, combined with concurrent milk MIR spectral data from routine milk recording, to train deep artificial neural networks to develop a prediction pipeline for bTB status. Such a tool would enable prediction of bTB status from milk

MIR spectral data alone and could be used as an early alert system as part of routine milk recording.

MATERIALS AND METHODS

Animals

Cow (n = 1,678,165) data were from national herds involved in routine milk recording with National Milk Records (**NMR**) and were distributed across GB. National Milk Records is the leading supplier of milk-recording services in the UK, processing a daily herd-level bulk-milk sample from 97% of UK farms as well as a monthly individual milk sample from 60% of the individual cows in the UK (National Milk Records, 2019). Since 2013, Scotland's Rural College has received spectral data daily, in addition to milk composition and pedigree information for cows from over 4,900 commercial farms across the UK 3 times per year. The majority of cows in this study were Holstein-Friesians (81%), followed by Belted Galloway (9%), Jersey (3%), Ayrshire (1%), Brown Swiss (0.8%), Swedish Red and White (0.8%), and Guernsey (0.7%). The data also included small numbers of other dairy breed and crosses (<3.7%).

Bovine Tuberculosis Data

Bovine tuberculosis data were made available by the Animal and Plant Health Agency and were collected via the GB national bTB testing program. These data provided information from over 40,500 confirmed and unconfirmed bTB herd breakdowns between October 2001 and January 2018, including breakdown start and end dates, breakdown duration, animal age at breakdown, SICCT skin-test date, lesion status, SICCT skin-test result, culture result, and slaughter status. Only data relating to dairy cows were considered in our study.

Cattle Movements Data

Data relating to cattle births, movements, and deaths were supplied by the British Cattle Movements Service. These data contained individual information relating to date, time and location of all births and deaths, as well as age at death. Additionally, processed data (i.e., calculated from the raw data) relating to any individual cattle movements were available with corresponding dates, locations (to and from), length of stays, distances traveled, location types (e.g., agricultural holding and slaughterhouse). These data were matched to concurrent bTB profiles of each cow in the study.

Mid-infrared Spectral Data

Milk Sampling and MIR Spectral Analysis. Milk sampling of individual cows occurred at 30-d intervals between January 2012 and August 2019 as part of a routine milk-recording service provided to farmers on a subscription basis. In addition to daily bulk-milk testing, NMR carried out MIR analysis of individual cow milk samples as part of their routine milk-recording services. For the present study, we focused on these routinely collected individual samples. Mid-infrared spectrometry of milk samples was carried out by National Milk Laboratories (Wolverhampton, UK), part of the NMR group, using FOSS FTIR spectrometers (FOSS Electric A/S, Hillerød, Denmark). The FOSS machines used an interferometer and the Fourier-transform infrared technique within the MIR region of wavelengths from 900 to 5,000 cm^{-1} to generate spectra (FOSS, 2016).

Pretreatment and Standardization of MIR Spectral Data. Following MIR analysis, a spectrum of 1,060 transmittance data points were generated; these data represented the absorption of infrared light through the milk sample. Before use in any analyses, the spectra were subject to several pretreatments. First, the transmittance data obtained from the spectrometer were converted to a linear absorbance scale by applying a $\log_{10}{}^{-0.5}$ transformation to the reciprocal of the transmittance (Soyeurt et al., 2011). Second, spectral data were standardized to account for drift incurred by collection of spectral data from different MIR instruments and across time (Grelet et al., 2015). Standardization was carried out using files supplied by the Walloon Agricultural Research Centre and following protocols developed within the InterReg/EU-funded project OptiMIR (Friedrichs et al., 2015). Standardization of the spectra as above had the added value of ensuring resultant-prediction tools could be applied to data streams from other machines throughout Europe that have adopted the same standardization procedure (Grelet et al., 2015), and that predictions could be compared across time because drift in the machines was accounted for.

Creation of Training and Testing Data Sets for Deep Learning

Definition of bTB Phenotype. The bTB phenotype was created for each cow using data relating to SICCT skin-test results, culture status, whether a cow was slaughtered, and whether any lesions were observed, all at the individual level. Information from each of these categories (where available) was combined

to create a binary phenotype; labeled 0 to represent nonresponders (i.e., healthy cows) and 1 to represent responders (i.e., bTB-affected cows). For example, if a skin test was inconclusive, but data indicated the cow was slaughtered and there was a positive observation of lesions, then this record was labeled as 1. Similarly, if a skin test suggested a nonresponder, but lesions were observed, then this record was also labeled as 1. Records were only ever labeled 0 when the skin-test result, combined with information relating to slaughter, culture, and lesions, did not indicate the presence of bTB.

Alignment of Spectral Data to bTB Profile. For each cow in the data set, bTB phenotype data (as described above) were matched to their concurrent milk MIR spectral data on sample date (i.e., the date of individual SICCT skin testing and individual milk sampling for bTB and spectral data, respectively). If no milk spectral data were collected on the same day as a SICCT skin test, then the milk spectra sample closest to skin-test date was used with a maximum tolerance of ± 15 d.

Data Preparation. To investigate the degree of accuracy of the bTB phenotype, as well as the effect of herd location, 3 distinct data sets were created. In all 3 data sets, responders were selected from confirmed bTB breakdown herds with nonresponders selected as follows: (1) nonresponders selected from herds with no confirmed responders, (2) nonresponders selected from the same heard breakdown as responders, and (3) nonresponders that eventually test positive for bTB, but the time between a negative (nonresponder) and positive (responder) result was greater than 183 d (i.e., a period of time sufficiently long enough to have observed multiple tests). Finally, data sets were randomly partitioned into training and validation sets for use in model development via deep learning. Data sets were partitioned such that approximately 80% of the data appeared in the training set with the remaining 20% in the validation set. Both training and validation data were balanced such that each set contained approximately equal numbers of reactors and nonreactors.

Deep Learning: Hardware and Software Requirements

To successfully use the power of deep learning in a timely manner, certain hardware and software requirements needed to be met. The full system specifications used in the present study are presented in Table 1 and summarized as follows: NVIDIA DGX Station personal AI supercomputer (NVIDIA Ltd., 2019) fitted with 4 NVIDIA Tesla V100 graphics processing units (**GPU**), Linux (Ubuntu) operating system, Python 3.5 Virtual Environment running within a Docker container, and PyTorch-GPU. PyTorch is an open source machine-learning library (released under the modified BSD license) developed by Facebook′s AI research group for use in research and development as well as production systems (Paszke et al., 2017). The GPU-enabled version of PyTorch offers enhanced processing speeds compared with the central processing unit version.

Table 1. System specifications of the deep-learning rig (building on a NVIDIA DGX Station[1])

Graphics processing units (GPU)	4× Tesla V100
TFLOPS[2] (mixed precision)	500
GPU memory	128 GB total system
NVIDIA tensor cores	2,560
NVIDIA CUDA[3] cores	20,480
Central processing unit	Intel Xeon E5–2698 v4 2.2 GHz (20-Core)
System memory	256 GB RDIMM DDR4
Storage	Data: 3X 1.92 TB SSD RAID 0
	OS: 1X 1.92 TB SSD
Network	Dual 10GBASE-T (RJ45)
Maximum power requirements	1,500 W
Operating system	Ubuntu Desktop Linux OS
Software	DGX Recommended GPU Driver
	CUDA Toolkit
	Docker
	Python 3.5
	Pytorch (GPU)
	Previously developed SRUC-EGENES Deep Learning pipelines[4]

[1]NVIDIA Ltd. (2019).

[2]TFLOPS = teraflops (i.e., the capability of a processor to calculate one trillion floating-point operations per second).

[3]CUDA = compute unified device architecture. Parallel computing architecture developed by NVIDIA to enable increases in computing performance by harnessing the power of the GPU to speed up demanding tasks.

[4]Brand et al. (2018).

Development of Prediction Tool

Repeated Observations. For the repeated observations on cows (i.e., only in the case of nonresponders), the only data used to train models were the 1,060 MIR wavelength values (i.e., features) with corresponding bTB status (i.e., labels). Deep-learning algorithms did not have access to any animal information, thus were unable to differentiate between multiple and single observations. Moreover, the majority of data (89%) were from single observations.

Data Synthesis. For supervised deep-learning tasks, an important requirement is a large quantity of balanced, labeled data (LeCun et al., 2015). In the case of bTB, the literature reports herd incidence of bTB of approximately 0.3 to 7.5% for low- and high-risk areas, respectively (Brotherstone et al., 2010). Furthermore, an incidence of approximately 4% was observed in the data available to the present study. The requirement for balanced labels (i.e., bTB-infected cows and healthy cows) meant that of the 250,000+ animal test dates available to us, we could only train approximately 20,000 due to the low number of bTB-positive records. To overcome this, we synthesized additional bTB-positive MIR spectra and investigated the effect of including these data during training. For the purposes of the present study, new data were synthesized using synthetic minority over sampling (**SMOTE**; Chawla et al., 2002) as well as the Adaptive Synthetic (**ADASYN**; He et al., 2008) sampling approach. Synthesized MIR data were only added to training sets, never to validation sets. Moreover, only bTB-positive MIR spectra were synthesized with labels balanced using real MIR spectral data from healthy cows.

Transfer Learning. Transfer learning is a machine-learning technique in which a pretrained (or learned) model, trained for a specific task, is repurposed for a new, different task (Goodfellow et al., 2016). This method enabled us to harness the knowledge and power of the vast amount of published research and development already available in the field of computer vision—the field with largest and most widely adopted use of deep learning. For our pretrained model, we opted to use DenseNet-161, a dense convolutional network where each layer in the network is connected to every other layer in a feed-forward fashion (Huang et al., 2017). This was made possible by converting individual spectral records into 53-pixel × 20-pixel greyscale images, as described below.

Creation of Images from MIR Spectral Wavelength Values. Mid-infrared spectral images were created by iterating through the data set, selecting an individual spectral record, and reshaping it from an array of size 1,060 × 1 to an array of size 53 × 20. Each of the reshaped arrays then had their wavelength values normalized to a value between 0 and 1 before finally multiplying each normalized wavelength by 255 to represent the wavelength values as grayscale pixels. Resulting arrays were then saved as individual PNG images (Figure 1).

Measures of Accuracy. To determine how well models performed, several metrics commonly used in machine and deep learning were calculated for resultant models. One of the most important of these metrics was loss, a value that ranges between 0 and $+\infty$ that is calculated by a specific loss function after each epoch during both training and validation (L_t and L_v, respectively, with $0 \leq L_{t,v}$). Loss functions are used to measure how wrong a model is (error) by comparing the predicted value, $\hat{y}$, with the actual value, y (LeCun et al., 2015). If the distance between $\hat{y}$ and y is large, then the loss will be high. Conversely, if the distance is small, then the loss will be low, thus providing an indication of model performance during training, as well any over- or under-fitting. Loss for models developed in the present study was calculated by pushing the final (output) layer through a softmax activation function (Equation 1); this ensured the output of each node was a probability between 0 and 1 before applying a log-loss function known as categorical cross entropy (Equation 2).

$$Softmax(y_i) = e^{yi}/\Sigma_j\, e^{yj}; \qquad [1]$$

$$Loss = -y\,\log(\hat{y}) - (1 - y)\,\log(1 - \hat{y}). \qquad [2]$$

Confusion matrices were created with a true positive (**TP**) recorded when the model correctly predicted the positive class (responders) and a true negative (**TN**) when the model correctly predicted the negative class (nonresponders). Similarly, a false positive (**FP**) was recorded when the model incorrectly predicted a nonresponder as a responder, and likewise, a false negative (**FN**) was recorded when the model incorrectly predicted a responder as a nonresponder. Ideally, one would want to minimize the number of FP and FN. False negatives were considered as extremely important because they would have serious ramifications in a live setting, resulting in potentially infected animals remaining in the herd. Total numbers of TP, TN, FP, and FN were then used to calculate additional metrics to determine model performance and included accuracy (**ACC**), precision, sensitivity (**TPR**), specificity (**TNR**), and the Matthews correlation coefficient (**MCC**).

Figure 1. Example of a spectral record represented as a grayscale image. Mid-infrared spectral images were created by reshaping spectral records from an array of size 1,060 × 1 to an array of size 53 × 20. Each wavelength value in the reshaped array was normalized (in the range 0–1), multiplying by 255 to represent the wavelength values as grayscale pixels, and saved as a PNG image. Image filenames were generated using label, animal, and sample information. These spectral images were then used as features in training the deep neural networks.

Accuracy was defined as the fraction of total predictions when the model was correct and was calculated as follows:

$$\mathrm{ACC} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}),$$

where $0 \leq \mathrm{ACC} \leq 1$.

Positive predictive value (**PPV**) is the probability that an individual with a positive test result is infected, and was defined as the proportion of positive predictions that were verified as correct; it was calculated as follows:

$$\mathrm{PPV} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}).$$

Thus, if a model produces no false positives, it would have a PPV of 1. Negative predictive value (**NPV**) is the probability that an individual with a negative test result is truly free from infection and was defined as the proportion of negative predictions that were verified as correct; it was calculated as follows:

$$\mathrm{NPV} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FN}).$$

Thus, if a model produces no false negatives it would have an NPV of 1.

Sensitivity (i.e., recall, or TP rate) was defined as the proportion of true positives the model identified correctly and was calculated as follows:

$$\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}).$$

Thus, if a model produces no false negatives, it would have a TPR of 1. Specificity (i.e., TN rate) was defined as the proportion of true negatives the model identified correctly and was calculated as follows:

$$\mathrm{TNR} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}).$$

Thus, if a model produces no false positives, it would have a TNR of 1.

Finally, the MCC (Matthews, 1975), a balanced measure of binary classifications used in machine learning and nondependent on which class is the positive class, was calculated via

$$\mathrm{MCC} = \frac{(\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}},$$

where $-1 \leq \mathrm{MCC} \leq 1$. It has been suggested that MCC is the most informative single-value measure in evaluating binary classification problems (Powers, 2007)

Table 2. Summary of the mid-infrared spectra-aligned bovine tuberculosis (bTB) baseline and training data sets[1]

Item	Baseline	Train[2]	Train[3]	Train[4]
Animal test dates	258,058	21,477	15,909	4,257
Cows	231,893	21,326	15,863	1,978
Herds	1,946	1,834	1,126	473
Herd breakdowns	2,936	2,210	1,326	1,109
Responders	8,591	8,591	6,771	1,978
Nonresponders	249,467	12,886	9,138	2,279

[1]Total numbers per data set of aligned records (animal test dates) are presented, broken down into total numbers of cows, herds, herd breakdowns, as well as the number of cows labeled as responders and nonresponders

[2]Nonresponders randomly selected from herds with no confirmed responders.

[3]Nonresponders from same herd-year-season as responders but never contracted bTB.

[4]Nonresponders eventually contract bTB (≥183 d between a positive and negative test).

because it considers the balance ratios of the confusion-matrix categories (Chicco, 2017).

RESULTS

Alignment of Spectral Data to bTB Profile

Alignment of bTB phenotypes with concurrent milk MIR spectral records produced a data set containing 259,957 animal test dates relating to 234,073 cows from 1,959 herds. There were 1,899 instances when the bTB phenotype could not be defined using the available data; these data were subsequently removed but retained for future use. Thus, the final data set for use in training models contained 258,058 animal test dates relating to 231,893 cows from 1,946 herds and concerned 2,936 distinct herd breakdowns. Regarding herd breakdowns, the majority (2,105) were confirmed breakdowns (i.e., officially tuberculosis-free–withdrawn, status), 809 were unconfirmed (i.e., officially tuberculosis-free–suspended status), and 22 were of unknown status. Descriptions of the data sets generated from these available data are summarized in Table 2.

Table 3. Measures of model performance resulting from training and validation

	Data set[1]		
Item	1	2	3
Epoch	750	750	278
Training loss	0.06	0.50	0.43
Validation loss	2.48	0.94	0.71
Accuracy	0.68	0.59	0.71
Positive predictive value	0.63	0.53	0.66
Negative predictive value	0.71	0.64	0.78
Sensitivity (true positive rate)	0.51	0.48	0.79
Specificity (true negative rate)	0.80	0.68	0.65
Matthews correlation coefficient	0.32	0.16	0.44

[1]Data set 1 = nonresponders randomly selected from herds with no confirmed responders. Data set 2 = nonresponders from same herd-year-season as responders but never contracted bovine tuberculosis. Data set 3 = nonresponders eventually contract bovine tuberculosis (≥183 d between a positive and negative test).

Development of the Prediction Tool

Results from training and validation are presented in Table 3. All models were trained in 2 stages; initially for 250 epochs for feature selection using the DenseNet161 pretrained model. The initial features passed to the DenseNet161 pretrained models were our grayscale MIR PNG images (described earlier); as such, the features selected by the model were not in the form of spectral wavelengths, but were in the form of higher-level features created as a result of passing the images through the CNN (Liu et al., 2016; Huang et al., 2017). Models were then trained for a further 500, 500, and 28 epochs for data sets 1, 2, and 3, respectively. The number of epochs required in both stages of training was determined by the inclusion of an early stopper in the code. Early stopping is a machine-learning method used to stop training when there is no improvement in model performance, thus minimizing over- and underfitting. In the case of our networks, validation loss was the metric that was monitored, with early stopping taking place when no improvement (i.e., minimize) was obtained over 25 iterations.

In general, model performance was greatest when developed using training data set 3 (0.71 ACC; 0.79 TPR; 0.65 TNR). Data set 1 showed the highest specificity (0.80), but also had a lower sensitivity (0.51) than the model developed using data set 3. Training using data set 2 resulted in the poorest performance (0.59 ACC; 0.48 TPR; 0.68 TNR). Data set 3 also required the least number of epochs to train, converging approximately 2.7 times faster. Comparing the MCC of the models developed using the 3 data sets (0.32, 0.16, and 0.44, for data sets 1, 2, and 3, respectively), we observed that data set 3 yielded the better model again. With all 3 MCC values less than 0.5, however, the MCC suggested that predicted label and the true label were only weakly to moderately correlated. This was further evidenced by the moderate PPV (0.63, 0.53, and 0.66, for data sets 1, 2, and 3, respectively) and NPV (0.71, 0.64, and 0.78, for data sets 1, 2, and 3, respectively) obtained.

Data Synthesis

Our investigations found that synthesizing data by applying SMOTE to real data returned improved results (higher ACC, lower L_v) compared with when ADASYN was applied; thus, SMOTE was chosen to

synthesize additional data for training our CNN. In all instances, the addition of synthesized data in training data sets (only real data were used for validation) resulted in increased model performance (Table 4) with observations of lower validation loss (0.46, 0.60, and 0.26 for data sets 1, 2, and 3, respectively) and a 1.32-, 1.32-, and 1.34-fold increase in accuracy for data sets 1, 2, and 3, respectively (0.90, 0.78, and 0.95 for data sets 1, 2, and 3, respectively). Improved sensitivity (0.85, 0.78, and 0.96 for data sets 1, 2, and 3, respectively), and specificity (0.93, 0.78, and 0.94 for data sets 1, 2, and 3, respectively) were also obtained when synthesized data were included in the training set. The MCC obtained were far more encouraging than those obtained previously (without synthesized data; Table 3), suggesting moderate (0.55 for data set 2) to strong (0.78 and 0.90 for data sets 1 and 3, respectively) correlations between predicted and true labels. Again, this was further evidenced by the strong PPV (0.89, 0.72, and 0.95, for data sets 1, 2, and 3, respectively) and NPV (0.90, 0.82, and 0.96, for data sets 1, 2, and 3, respectively) obtained. The results from data set 3 signified the model was able to successfully distinguish between spectra from bTB-positive and bTB-negative cows, with a high probability that those flagged as bTB-infected and noninfected were infected and free from infection, respectively.

DISCUSSION

The present study developed a pipeline for the prediction of bTB status in dairy cows by applying state-of-the-art deep-learning techniques to their milk MIR spectral profiles. The prospect of using routinely collected milk samples for the early identification of bTB-infected cows represents an innovative, low-cost and, importantly, noninvasive tool that has the potential to contribute substantially in the push to eradicate bTB in England, Wales, and the wider UK. Such a tool would not only complement the current control measures (e.g., intradermal skin test, interferon-gamma assay), but also facilitate the rapid and seamless delivery of vital information to farmers, allowing them to make fast and informed management decisions that would significantly increase the health and welfare of their animals in addition to reducing costs to the farm, government, and taxpayer. If such a form of surveillance were to become approved, certain contingencies would have to be put in place; for example, Defra would need to be informed in the first instance to stop the illegal movement of alerted animals.

Table 4. Measures of model performance resulting from training and validation where training data sets contained both real and synthesized mid-infrared spectral data

	Data set[1]		
Item	1	2	3
Epoch	750	750	295
Training loss	0.22	0.37	0.23
Validation loss	0.46	0.60	0.26
Accuracy	0.90	0.78	0.95
Positive predictive value	0.89	0.72	0.95
Negative predictive value	0.90	0.82	0.96
Sensitivity (true positive rate)	0.85	0.78	0.96
Specificity (true negative rate)	0.93	0.78	0.94
Matthews correlation coefficient	0.78	0.55	0.90

[1]Data set 1 = nonresponders randomly selected from herds with no confirmed responders. Data set 2 = nonresponders from same herd-year-season as responders but never contracted bovine tuberculosis. Data set 3 = nonresponders eventually contract bovine tuberculosis (≥183 d between a positive and negative test).

Harnessing the Power of Big Data and Artificial Intelligence

The standard method of calibrating milk MIR spectral data using matched phenotypes by partial least squares regression has delivered several successful quantitative analysis tools as highlighted (De Marchi et al., 2014). In the case of phenotypes represented by discrete data (e.g., categorical and binary) the usual methods for developing prediction equations have proved less efficient and resulted in lower accuracy predictions (Toledo-Alvarado et al., 2018; Delhez et al., 2020). Hence, there is a requirement for alternative and novel mathematical and statistical techniques to better use milk MIR spectra, a requirement we believe we have shown can be met using machine learning.

As previously mentioned, deep learning is a branch of the larger field of machine learning that uses algorithms that are better able to make use of today's ever-growing repositories of data and advances in computer technology (Bengio, 2009; Deng et al., 2009; Krizhevsky et al., 2012; LeCun et al., 2015). Deep learning is now being used to develop solutions to problems in a variety of research fields from medicine (e.g., diagnosing unknown skin lesions; Kawahara et al., 2016) to transportation (e.g., self-driving vehicles; Martinez et al., 2017). Further examples of deep learning can be found powering the mobile phone in your pocket and the smart technologies in your home. In the agricultural and animal sciences, uptake of deep-learning techniques has been slow (Howard, 2018). Recently, however, our group applied a deep CNN to MIR spectra-matched pregnancy data and discovered such algorithms significantly improved the prediction accuracy for pregnancy status in dairy cows, a binary phenotype (Brand et al., 2018).

Deep-learning tasks are known to require large volumes of data to successfully train a network. Moreover, for supervised learning problems, such as in the present study, there is an additional requirement that data labels must be more or less equally distributed (LeCun et al., 2015; Goodfellow et al., 2016). When the incidence of bTb is low (~4% in our data), one label dominates the data. Training on such a data set would result in an immensely inaccurate model, and the simple approach of under-sampling would greatly reduce the amount of data available for training. To overcome these challenges, we adopted 2 separate approaches, one to increase the size of our training data set (data synthesis), and another to lessen the effect of data size (transfer learning).

Data synthesis is a technique commonly applied in machine learning for many different purposes, from creating naïve, clean data for training models (Mikołajczyk and Grochowski, 2018) to overcoming privacy or legal issues when working with financial or medical data (Choi et al., 2017). To synthesize data for our purpose, we investigated 2 popular and widely used techniques, SMOTE and ADASYN. Both of these techniques use a *k*-nearest neighbors approach to synthesize new data within the body of available data by randomly selecting a minority instance, A, finding its *k*-nearest neighbors, and then drawing a line segment in the feature space between A and a random neighbor. Synthetic instances are then generated on the line (Chawla et al., 2002; He, 2011). The ADASYN technique modifies SMOTE slightly to synthesize more instances in regions of the feature space where minority instances are sparse, and fewer (or none) where minority instances are dense (He et al., 2008). There are many other approaches available to synthesize data, some of which are more advanced (themselves underpinned by deep learning), such as generative adversarial networks. The generative adversarial network uses 2 neural networks that are pitted against one another: a generative network which generates synthetic examples and a discriminative network which evaluates them to determine if they are real or synthetic. The aim of the generative network is to trick the discriminative network into labeling a synthetic instance as real (Goodfellow et al., 2011).

Another approach to enable the training of networks with less data available is that of transfer learning. In this approach, a model developed for one task is repurposed as a starting point and fine-tuned to develop a model for a different task. Developing neural-network models using deep learning requires high levels of resource in the form of both compute and time. As such, using a pretrained model as a starting point, and subsequently fine-tuning it for a specific problem or task, can provide massive gains (Pan and Yang, 2010; Shin et al., 2016; Yang et al., 2020).

Transfer learning combined with data synthesis can provide an effective enabling method for carrying out deep-learning tasks when the underlying data set size is on the smaller side, as evidenced by the present study. We were able to train a model to predict the bTB phenotype (as defined above) with 95% accuracy with a strong correlation between predicted and true labels (MCC = 0.90).

The current SICCT skin test has a high specificity (99.98%), indicating a high confidence in results where cows fail the test. Conversely, the sensitivity is not as high (ranges between 52–100%; average of 80%) indicating that not all cows that pass the test are truly bTB-free (i.e., some bTB-infected individuals are missed; de la Rua-Domenech et al., 2006). The current gamma interferon test, a more expensive test used alongside the SICCT test, is known to have a higher sensitivity than the SICCT test (~85–90%), but a lower specificity of 96.6% (Ryan et al., 2000; de la Rua-Domenech et al., 2006). Although our proposed tool has a slightly lower specificity that the SICCT test (96%), it is approximately equal to that of the gamma interferon test. Furthermore, we obtained a higher sensitivity than both of the current testing methods (94%), implying less false negatives will find return to the herd to infect other susceptible individuals.

The present study reinforces the utility of a deep-learning approach to calibrate MIR spectra to predict economically important and hard-to-record phenotypes. We believe our study to be the first of its kind to use deep learning to calibrate MIR spectra for phenotype prediction. Furthermore, we believe this to be the first study to use MIR spectral data to predict bTB, as well as the first to predict a contagious disease phenotype in general. The success of the prediction opens up the possibility to calibrate MIR spectra for other economically important diseases such as Paratuberculosis (Johne's disease), a chronic and contagious enteritis of ruminants caused by the bacterium *Mycobacterium avium* ssp. *Paratuberculosis*.

Existing bTB Control Measures and Possible Applications of the MIR-based Tool

The current bTB control strategy applied throughout GB is a combination of statutory and voluntary measures that are dependent on the perceived level of bTB risk in the area. The control measures applied to all areas, regardless of risk, can be split into 4 categories: surveillance, breakdown management, risk from badgers, and other disease prevention (DEFRA, 2014).

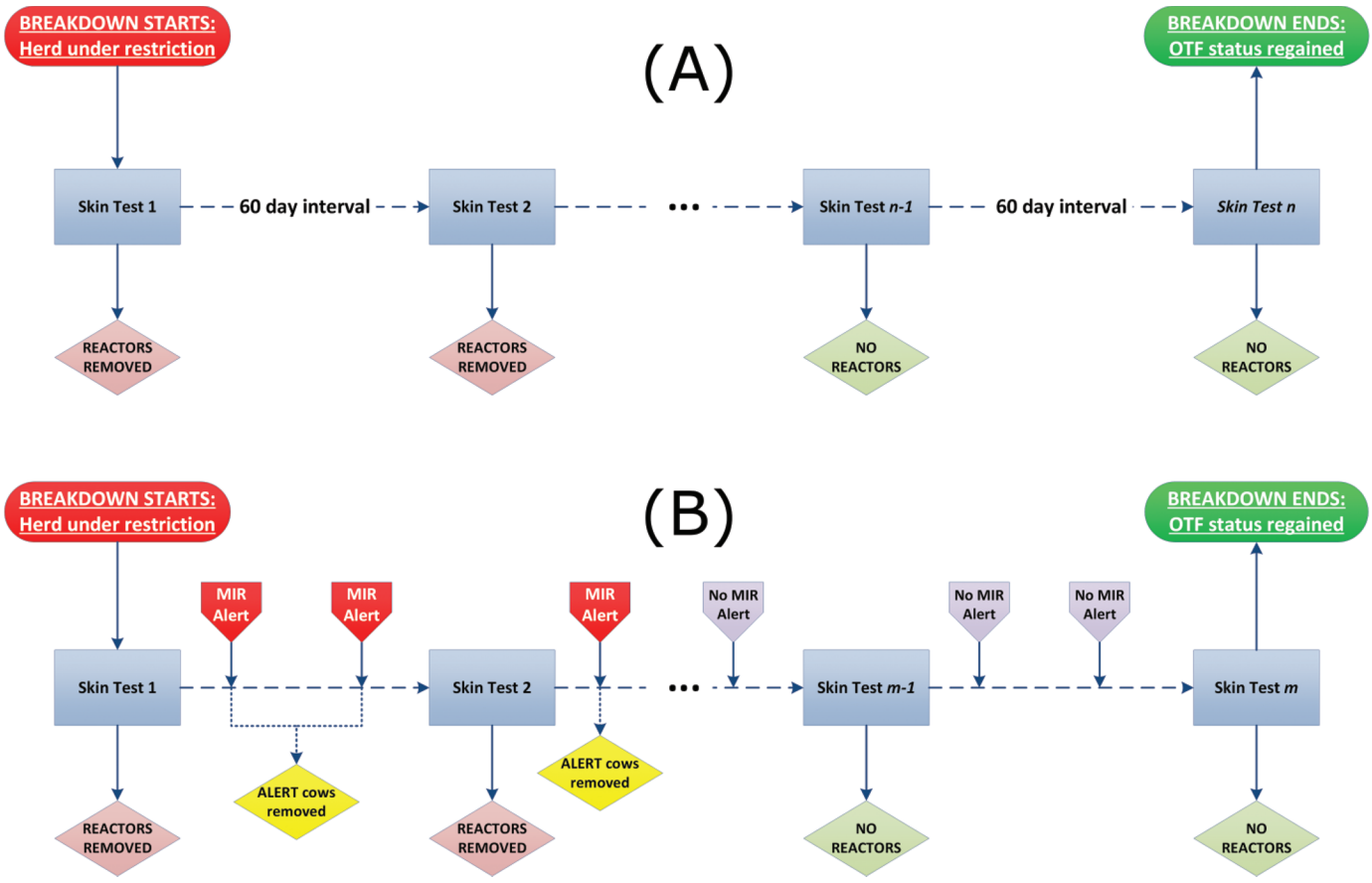

Figure 2. Schema for mid-infrared (MIR) alerts for bovine tuberculosis (bTB). (A) Flow diagram showing current bTB cattle control measures and (B) hypothesized control measures resulting from the MIR-based tool. Once bTB is disclosed, the herd is put under restriction and subjected to skin tests every 60 d until 2 sequential test periods result in no reactors. The total length of a breakdown is therefore $60 \times (n - 1)$ days, where n is the number of tests ($n > 2$). Due to the nature of bTB (infectious, chronic, and slowly progressive), one breakdown has the potential to last for months or even years. (B) shows the potential of the MIR prediction pipeline to alert the farmer to cows that will fail the skin test, allowing these "alerted" cows to be removed from the herd earlier, and thus reducing the spread of bTB in the herd. This offers the potential to significantly reduce the number of days for a restricted herd to regain officially tuberculosis-free (OTF) status [e.g., from $60 \times (n - 1)$ days to $60 \times (m - 1)$ days, where m is the number of tests ($m > 2$) and $m < n$].

Our proposed MIR-based tool would complement both the surveillance and breakdown-management areas of the current control strategy as discussed below.

Surveillance. At present, key measures include on farm statutory testing as well as carcass testing at the abattoir. Results from the present study highlight the value of a MIR spectra-based alert of potential bTB infection within a herd, specifically enabling the farmer (or a veterinarian) to identify and isolate (or cull) animals ahead of routine testing both on farm and at the abattoir. This would be especially beneficial in the case of herds with "officially tuberculosis free" status with no history of bTB outbreaks, allowing farmers to monitor their herd through routine milk recording and minimize the length of a breakdown if bTB is subsequently discovered. Additionally, when alerts arise from milk MIR (animals likely to be exposed above a minimum threshold of accuracy), a herd test may be triggered, allowing the farm to officially identify and isolate or quarantine potential reactors.

Once removed at an earlier stage, infected animals would have a reduced opportunity to infect other animals (or other wildlife reservoirs), thus leading to a reduction in the overall level of herd infectivity. This may eventually reduce the basic reproductive number (R_0) to a level such that other interventions have a greater effect. The R_0 of an infection is defined as the average number of secondary infections produced by an infected individual in a completely susceptible host population, and determines whether or not the infection can persist (Anderson and May, 1991).

Breakdown Management. For herds under (or on the onset of) restriction, the proposed tool has the potential to significantly reduce the length of the break-

down (Figure 2A). At present, once bTB is disclosed, the herd is put under restriction and subjected to skin tests every 60 d until 2 successive test periods result in no reactors. The total length of a breakdown can therefore be calculated as $60 \times (n - 1)$ days (where n = number of skin tests, and $n > 2$), and due to the infectious, chronic, and slowly progressive nature of bTB, 1 breakdown has the potential to last for months, years, or even decades.

This is where early identification of infected animals would be advantageous. Alerting the farmer to cows that will fail the next skin test allows them to be removed from the herd, reducing the spread of bTB. This offers the potential to significantly reduce the length of restriction [e.g., from $60 \times (n - 1)$ days to $60 \times (m - 1)$ days, where m is the number of tests ($m > 2$) and $m < n$ (Figure 2B)]. Moreover, for farms already involved in routine milk recording, such a system would require no additional labor or changes in management.

CONCLUSIONS

Deep learning, underpinned by convolutional neural networks, has provided a promising method to calibrate milk MIR spectral data to predict bTB status of individual dairy cows. The models developed were able to successfully alert which cows would be expected to fail the SICCT skin test, with an accuracy of 95% and a corresponding sensitivity and specificity of 0.96 and 0.94, respectively. Moreover, predictions were strongly correlated with true values (MCC = 0.90). The automated prediction of bTB status at routine milk recording could provide farmers with a robust tool that enables them to make early management decisions on potential reactor cows. The tool would have the added benefit of providing an effective enabling service, giving farmers the opportunity to be more engaged with bTB testing, as well as the ability to take ownership of the health of their herd. Such a tool would also provide the government with an additional mechanism to have an immediate and enduring effect on the prevalence of bTB in UK dairy herds.

ACKNOWLEDGMENTS

This work was supported by a Biotechnology and Biological Sciences Research Council (BBSRC) Industrial Partnership Award (grant no. BB/S009396/1) awarded to MC and carried out in partnership with National Milk Records (NMR, Chippenham, UK). Milk spectral data were provided by NMR and authors gratefully acknowledge collaboration with NMR (Martin Busfield, Eamon Watson, and Andy Warne). Herd breakdown data were provided by the Animal and Plant Health Agency (APHA, Addlestone, UK) and British Cattle Movement Service (BCMS, Workington, UK). We thank Ian Archibald (SRUC, Edinburgh, Scotland) for managing the data and assisting with extraction. The authors confirm that they have no conflicts of interest.

REFERENCES

Allen, A. R., G. Minozzi, E. J. Glass, R. A. Skuce, S. W. J. McDowell, J. A. Woolliams, and S. C. Bishop. 2010. Bovine tuberculosis: the genetic basis of host susceptibility. Proc. Biol. Sci. 277:2737–2745. https://doi.org/10.1098/rspb.2010.0830.

Anderson, R. M., and R. M. May. 1991. Infectious Diseases of Humans: Dynamics and Control. Oxford University Press, Oxford, UK.

Banos, G., M. Winters, R. Mrode, A. P. Mitchell, S. C. Bishop, J. A. Woolliams, and M. P. Coffey. 2017. Genetic evaluation for bovine tuberculosis resistance in dairy cattle. J. Dairy Sci. 100:1272–1281. https://doi.org/10.3168/jds.2016-11897.

Bengio, Y. 2009. Learning deep architectures for AI. Found. Trends. in Mach. Learn. 2:1–55. https://doi.org/http://doi.org/10.1561/2200000006.

Bermingham, M. L., S. J. More, M. Good, A. R. Cromie, I. M. Higgins, S. Brotherstone, and D. P. Berry. 2009. Genetics of tuberculosis in Irish Holstein-Friesian dairy herds. J. Dairy Sci. 92:3447–3456. https://doi.org/10.3168/jds.2008-1848.

Berry, D. P., S. McParland, C. Bastin, E. Wall, N. Gengler, and H. Soyeurt. 2013. Phenotyping of robustness and milk quality. Adv. Anim. Biosci. 4:600–605. https://doi.org/10.1017/S2040470013000150.

Brand, W., A. T. Wells, and M. P. Coffey. 2018. Predicting pregnancy status from mid-infrared spectroscopy in dairy cow milk using deep learning. J. Dairy Sci. 101(Suppl. 2):34.

Brotherstone, S., I. M. S. White, M. P. Coffey, S. H. Downs, A. P. Mitchell, R. S. Clifton-Hadley, S. J. More, M. Good, and J. A. Woolliams. 2010. Evidence of genetic resistance of cattle to infection with *Mycobacterium bovis*. J. Dairy Sci. 93:1234–1242. https://doi.org/10.3168/jds.2009-2609.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. J. Artif. Intell. Res. 16:321–357. https://doi.org/10.1613/jair.953.

Chicco, D. 2017. Ten quick tips for machine learning in computational biology. BioData Min. 10:35. https://doi.org/10.1186/s13040-017-0155-3.

Choi, E., S. Biswal, B. Malin, J. Duke, W.F. Stewart, and J. Sun. 2017. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. Accessed Jul. 2020. https://www.researchgate.net/publication/315456455_Generating_Multi-label_Discrete_Patient_Records_using_Generative_Adversarial_Networks.

Crimes, D., and G. Enticott. 2019. Assessing the social and psychological impacts of endemic animal disease amongst farmers. Front. Vet. Sci. 6:342. https://doi.org/10.3389/fvets.2019.00342.

AHDB Dairy. 2016. TB Advantage - The Genetics of BTB. Accessed Apr. 25, 2018. https://dairy.ahdb.org.uk/technical-information/breeding-genetics/tb-advantage.

de la Rua-Domenech, R., A. T. Goodchild, H. M. Vordermeier, R. G. Hewinson, K. H. Christiansen, and R. S. Clifton-Hadley. 2006. Ante mortem diagnosis of tuberculosis in cattle: A review of the tuberculin tests, γ-interferon assay and other ancillary diagnostic techniques. Res. Vet. Sci. 81:190–210. https://doi.org/10.1016/j.rvsc.2005.11.005.

De Marchi, M., V. Toffanin, M. Cassandro, and M. Penasa. 2014. Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. J. Dairy Sci. 97:1171–1186. https://doi.org/10.3168/jds.2013-6799.

DEFRA. 2014. The Strategy for achieving Officially Bovine Tuberculosis Free status for England. Accessed Jul. 2020. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/300447/pb14088-bovine-tb-strategy-140328.pdf.

Dehareng, F., C. Delfosse, E. Froidmont, H. Soyeurt, C. Martin, N. Gengler, A. Vanlierde, and P. Dardenne. 2012. Potential use of milk mid-infrared spectra to predict individual methane emission of dairy cows. Animal 6:1694–1701. https://doi.org/10.1017/S1751731112000456.

Delhez, P., P. N. Ho, N. Gengler, H. Soyeurt, and J. E. Pryce. 2020. Diagnosing the pregnancy status of dairy cows: How useful is milk mid-infrared spectroscopy? J. Dairy Sci. 103:3264–3274. https://doi.org/10.3168/jds.2019-17473.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li. 2009. ImageNet: A large-scale hierarchical image database. Pages 248–255 in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Miami, FL. IEEE, New York, NY.

Department for Environment, Food and Rural Affairs. 2018. Quarterly Publication of National Statistics on the Incidence and Prevalence of Tuberculosis (TB) in Cattle in Great Britain – to End December 2017. Accessed Apr. 25, 2018. https://www.gov.uk/government/statistics/incidence-of-tuberculosis-tb-in-cattle-in-great-britain.

FarmingUK. 2018. Stress and Depression Common Causes of Ill Health in Farming. Accessed Nov. 26, 2019. https://www.farminguk.com/news/stress-and-depression-common-causes-of-ill-health-in-farming_50623.html.

FOSS. 2016. FTIR Analysis of Food and Agric. Products. Accessed Jun. 7, 2016. http://www.foss.dk/.

Friedrichs, P., C. Bastin, F. Dehareng, B. Wickham, and X. Massart. 2015. Final OptiMIR scientific and expert meeting: From milk analysis to advisory tools (Palais des Congrès, Namur, Belgium, 16–17 April 2015). Biotechnol. Agron. Soc. Environ. 19:97–124.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. Deep Learning. MIT Press, Cambridge, MA.

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2011. Generative adversarial nets. Page 085201 in Advances in Neural Information Processing Systems. Elsevier, the Netherlands.

Grelet, C., C. Bastin, M. Gelé, J.-B. Davière, M. Johan, A. Werner, R. Reding, J. A. Fernandez Pierna, F. G. Colinet, P. Dardenne, N. Gengler, H. Soyeurt, and F. Dehareng. 2016. Development of Fourier transform mid-infrared calibrations to predict acetone, β-hydroxybutyrate, and citrate contents in bovine milk through a European dairy network. J. Dairy Sci. 99:4816–4825. https://doi.org/10.3168/jds.2015-10477.

Grelet, C., J. A. Fernández Pierna, P. Dardenne, V. Baeten, and F. Dehareng. 2015. Standardization of milk mid-infrared spectra from a European dairy network. J. Dairy Sci. 98:2150–2160. https://doi.org/10.3168/jds.2014-8764.

He, H. 2011. Imbalanced Learning. H. He and Y. Ma, ed. John Wiley & Sons Inc., Hoboken, NJ.

He, H., Y. Bai, E. A. Garcia, and S. Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. Pages 1322–1328 in 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong. IEEE, New York, NY.

Howard, J. 2018. Deep Learning: The tech that's changing everything, except animal breeding and genetics [plenary address]. 11th World Congress on Genetics Applied to Livestock Production, Auckland, New Zealand. https://icarinterbullwcgalp.zerista.com/event/member/453201.

Huang, G., Z. Liu, L. van der Maaten, and K. Q. Weinberger. 2017. Densely Connected Convolutional Networks. Pages 2261–2269 in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI. IEEE, New York, NY.

Humblet, M. F., M. L. Boschiroli, and C. Saegerman. 2009. Classification of worldwide bovine tuberculosis risk factors in cattle: A stratified approach. Vet. Res. 40:50. https://doi.org/10.1051/vetres/2009033.

Kawahara, J., A. Bentaieb, and G. Hamarneh. 2016. Deep features to classify skin lesions. Pages 1397–1400 in Proc. Int. Symp. Biomed. Imaging, Prague, Czech Republic. IEEE, New York, NY. https://doi.org/10.1109/ISBI.2016.7493528.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25:1–9.

Lainé, A., H. Bel Mabrouk, L. Dale, C. Bastin, and N. Gengler. 2014. How to use mid-infrared spectral information from milk recording system to detect the pregnancy status of dairy cows. Commun. Agric. Appl. Biol. Sci. 79:33–38.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. Nature 521:436–444. https://doi.org/10.1038/nature14539.

Liu, Z., J. Gao, G. Yang, H. Zhang, and Y. He. 2016. Localization and classification of paddy field pests using a saliency map and deep convolutional neural network. Sci. Rep. 6:20410. https://doi.org/10.1038/srep20410.

Martinez, M., C. Sitawarin, K. Finch, L. Meincke, A. Yablonski, and A. Kornhauser. 2017. Beyond Grand Theft Auto V for Training, Testing and Enhancing Deep Learning in Self Driving Cars. Accessed Jul. 2020. https://arxiv.org/abs/1712.01397.

Matthews, B.W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta. Bioenerg. 405:442–451.

McParland, S., G. Banos, E. Wall, M. P. Coffey, H. Soyeurt, R. F. Veerkamp, and D. P. Berry. 2011. The use of mid-infrared spectrometry to predict body energy status of Holstein cows. J. Dairy Sci. 94:3651–3661. https://doi.org/10.3168/jds.2010-3965.

Mikołajczyk, A., and M. Grochowski. 2018. Data augmentation for improving deep learning in image classification problem. Pages 117–122 in 2018 International Interdisciplinary PhD Workshop (IIPhDW), Swinoujście, Poland. IEEE, New York, NY. 10.1109/IIPHDW.2018.8388338.

National Milk Records. 2019. History of NMR - National Milk Records. Accessed Oct. 18, 2019. https://www.nmr.co.uk/company/history.

NVIDIA Ltd. 2019. NVIDIA DGX Station: AI Workstation for Data Science Teams. Accessed Nov. 27, 2019. https://www.nvidia.com/en-gb/data-center/dgx-station.

Olea-Popelka, F., A. Muwonge, A. Perera, A. S. Dean, E. Mumford, E. Erlacher-Vindel, S. Forcella, B. J. Silk, L. Ditiu, A. El Idrissi, M. Raviglione, O. Cosivi, P. LoBue, and P. I. Fujiwara. 2017. Zoonotic tuberculosis in human beings caused by *Mycobacterium bovis*—A call for action. Lancet Infect. Dis. 17:e21–e25. https://doi.org/10.1016/S1473-3099(16)30139-6.

Pan, S. J., and Q. Yang. 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22:1345–1359. https://doi.org/10.1109/TKDE.2009.191.

Parry, J., R. Lindsey, and R. Taylor. 2005. Farmers, farm workers and work-related stress. Accessed Jul. 2020. https://www.researchgate.net/publication/277843723_Farmers_farm_workers_and_work-related_stress.

Paszke, A., S. Gross, S. Chintala, G. Chanan, E. Yang, et al. 2017. Automatic Differentiation in PyTorch. Accessed Aug. 9, 2019. https://openreview.net/pdf?id=BJJsrmfCZ.

Pollock, J. M., and S. D. Neill. 2002. *Mycobacterium bovis* infection and tuberculosis in cattle. Vet. J. 163:115–127. https://doi.org/10.1053/tvjl.2001.0655.

Powers, D. M. W. 2007. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation (spie-07–001). Tech. Rep. Adelaide, Australia.

Ryan, T. J., B. M. Buddle, and G. W. De Lisle. 2000. An evaluation of the gamma interferon test for detecting bovine tuberculosis in cattle 8 to 28 days after tuberculin skin testing. Res. Vet. Sci. 69:57–61. https://doi.org/10.1053/rvsc.2000.0386.

Shin, H. C., H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. 2016. Deep convolutional neural networks for computer-aided detection: cnn architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging 35:1285–1298. https://doi.org/10.1109/TMI.2016.2528162.

Smith, S. L., S. J. Denholm, M. P. Coffey, and E. Wall. 2019. Energy profiling of dairy cows from routine milk mid-infrared analysis.

J. Dairy Sci. 102:11169–11179. https://doi.org/10.3168/jds.2018-16112.

Soyeurt, H., C. Bastin, F. G. Colinet, V. M. R. Arnould, D. P. Berry, E. Wall, F. Dehareng, H. N. Nguyen, P. Dardenne, J. Schefers, J. Vandenplas, K. Weigel, M. Coffey, L. Théron, J. Detilleux, E. Reding, N. Gengler, and S. McParland. 2012. Mid-infrared prediction of lactoferrin content in bovine milk: Potential indicator of mastitis. Animal 6:1830–1838. https://doi.org/10.1017/S1751731112000791.

Soyeurt, H., P. Dardenne, F. Dehareng, G. Lognay, D. Veselko, M. Marlier, C. Bertozzi, P. Mayeres, and N. Gengler. 2006. Estimating fatty acid content in cow milk using mid-infrared spectrometry. J. Dairy Sci. 89:3690–3695. https://doi.org/10.3168/jds.S0022-0302(06)72409-2.

Soyeurt, H., F. Dehareng, N. Gengler, S. McParland, E. Wall, D. P. Berry, M. P. Coffey, and P. Dardenne. 2011. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. J. Dairy Sci. 94:1657–1667. https://doi.org/10.3168/jds.2010-3408.

Toledo-Alvarado, H., A. I. Vazquez, G. de los Campos, R. J. Tempelman, G. Bittante, and A. Cecchinato. 2018. Diagnosing pregnancy status using infrared spectra and milk composition in dairy cows. J. Dairy Sci. 101:2496–2505. https://doi.org/10.3168/jds.2017-13647.

Tsairidou, S., J. A. Woolliams, A. R. Allen, R. A. Skuce, S. H. McBride, D. M. Wright, M. L. Bermingham, R. Pong-Wong, O. Matika, S. W. J. McDowell, E. J. Glass, and S. C. Bishop. 2014. Genomic prediction for tuberculosis resistance in dairy cattle. PLoS One 9:e96728. https://doi.org/10.1371/journal.pone.0096728.

Veerkamp, R. F., L. Kaal, Y. de Haas, and J. D. Oldham. 2013. Breeding for robust cows that produce healthier milk: RobustMilk. Adv. Anim. Biosci. 4:594–599. https://doi.org/10.1017/S2040470013000149.

Wallén, S. E., E. Prestløkken, T. H. E. Meuwissen, S. McParland, and D. P. Berry. 2018. Milk mid-infrared spectral data as a tool to predict feed intake in lactating Norwegian Red dairy cows. J. Dairy Sci. 101:6232–6243. https://doi.org/10.3168/jds.2017-13874.

Yang, Q., Y. Zhang, W. Dai, and S. J. Pan. 2020. Transfer Learning. Cambridge University Press, Cambridge, UK.

ORCIDS

S. J. Denholm https://orcid.org/0000-0002-6291-8269
W. Brand https://orcid.org/0000-0002-5523-7510
A. P. Mitchell https://orcid.org/0000-0002-4059-0362
A. T. Wells https://orcid.org/0000-0001-8580-0426
E. Wall https://orcid.org/0000-0002-7072-5758
M. P. Coffey https://orcid.org/0000-0003-4890-6218